MIT Technology Review Insights

Produced in partnership with

databricks × dbt Labs

Emerging generative AI applications are pressuring data and analytics engineering teams to ship trusted data faster—how are data practitioners responding?

The data practitioner for the AI era

Preface

"The data practitioner for the AI era" is an MIT Technology Review Insights report sponsored by Databricks and dbt Labs. This report draws on in-depth interviews with executives and data leaders at data-focused organizations, conducted between September and December 2023.

Adam Green was the author of the report, Teresa Elsey was the editor, and Nicola Crepaldi was the publisher. The research is editorially independent, and the views expressed are those of MIT Technology Review Insights.

We would like to thank the following executives for their time and insights:

Drew Banin, Cofounder, dbt Labs

Venkatesh Kumar, Director of Data Integrations, Apixio

Niklas Nordansjö, Product Manager, Tibber

David Samet, Director of Technology, Fabuwood

Valdar Tammik, Head of Data, Starship Technologies

Reynold Xin, Cofounder and Chief Architect, Databricks

Srinivas Yadlapati, Head of Data, StockX

CONTENTS

01

Executive summary

The rise of generative AI, coupled with the rapid adoption and democratization of AI across industries this decade, has emphasized the singular importance of data. Managing data effectively has become critical to this era of business – making data practitioners, including data engineers, analytics engineers, and ML engineers, key figures in the data and AI revolution.

Organizations that fail to use their own data will fall behind competitors that do and miss out on opportunities to uncover new value for themselves and their customers. As the quantity and complexity of data grows, so do its challenges, forcing organizations to adopt new data tools and infrastructure which, in turn, change the roles and mandate of the technology workforce.

Data practitioners are among those whose roles are experiencing the most significant change, as organizations expand their responsibilities. Rather than working in a siloed data team, data engineers are now developing platforms and tools whose design improves data visibility and transparency for employees across the organization, including analytics engineers, data scientists, data analysts, machine learning engineers, and business stakeholders.

This report explores, through a series of interviews with expert data practitioners, key shifts in data engineering, the evolving skill set required of data practitioners, options for data infrastructure and tooling to support AI, and data challenges and opportunities emerging in parallel with generative AI. The report's key findings include the following.

- **The foundational importance of data is creating new demands on data practitioners.** As the rise of AI demonstrates the business importance of data more clearly than ever, data practitioners are encountering new data challenges, increasing data complexity, evolving team structures, and emerging tools and technologies – as well as establishing newfound organizational importance.

- **Data practitioners are getting closer to the business, and the business closer to the data.** The pressure to create value from data has led executives to invest more substantially in data-related functions. Data practitioners are being asked to expand their knowledge of the business, engage more deeply with business units, and support the use of data in the organization, while functional teams are finding they require their own internal data expertise to leverage their data.

- **The data and AI strategy has become a key part of the business strategy.** Business leaders need to invest in their data and AI strategy – including making important decisions about the data team's organizational structure, data platform and architecture, and data governance – because every business's key differentiator will increasingly be its data.

- **Data practitioners will shape how generative AI is deployed in the enterprise.** The key considerations for generative AI deployment – producing high-quality results, preventing bias and hallucinations, establishing governance, designing data workflows, ensuring regulatory compliance – are the province of data practitioners, giving them outsize influence on how this powerful technology will be put to work.

02

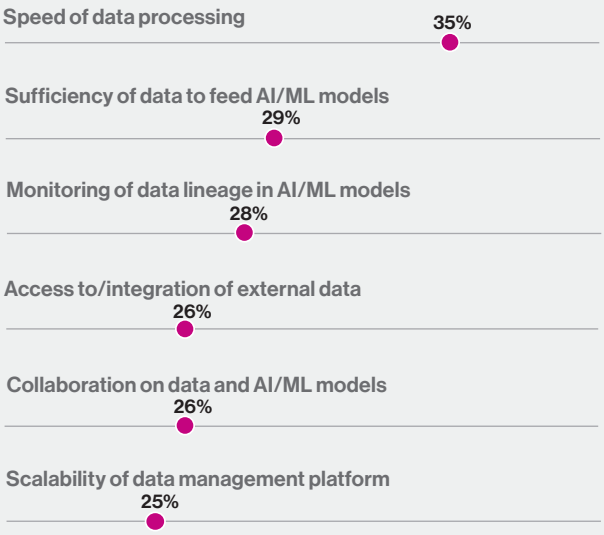
The changing role of data in the AI era

Even before the advent of generative AI, the scale and complexity of data were increasing exponentially. Now, as generative AI is triggering companies in all industries to double down on their technology investments, and as advanced use cases for data become ubiquitous in every sector, organizations are focusing even more on their data teams and tools.

The degree to which data limitations hamper an organization's AI ambitions is well-understood in the C-suite. In our 2022 survey, 72% of executives agreed that data problems were the most likely issue to jeopardize their AI/ML goals.[1] They cited speed of data processing (35%), sufficiency of data to feed AI/ML models (29%), and monitoring of data lineage in AI/ML models (28%) among the top data challenges requiring improvements. (See Figure 1.)

Emerging data platforms and tools are improving data visibility and transparency for employees across the organization, not just those in traditionally data-centric roles, such as data engineers, data scientists, and ML experts. Simultaneously, data engineering has emerged as a key function that builds, supports, and manages these data-driven systems.

Figure 1: Data challenges limit AI aspirations
Executives identified the aspects of their data strategy most in need of improvement to enable the organization to achieve future AI goals. (Respondents could select more than one answer.)



Source: "Becoming an AI-driven enterprise," MIT Technology Review Insights, September 20, 2022[2]

Data engineers' job descriptions have, historically, focused on service-oriented tasks, responding to the ad hoc needs of the business alongside managing infrastructure and systems. But expectations are changing. As data is increasingly understood as a key company asset, data practitioners have taken a role central to the business strategy, overseeing all aspects of the data ecosystem, from architecture to analytics to governance.

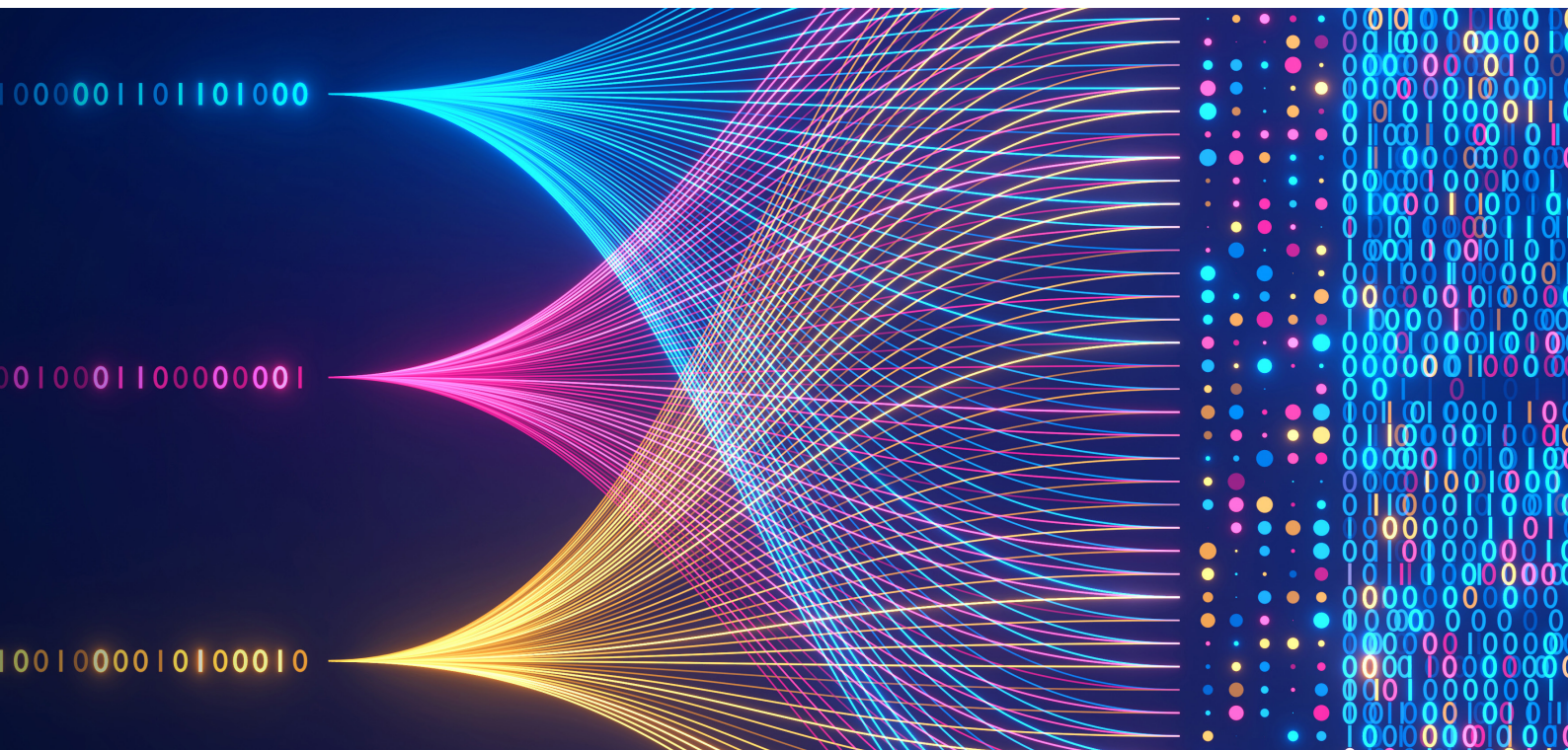
The new data engineer

The role of data engineer emerged in response to the growth in data volume and complexity in the 2010s. The development of data lakes enabled companies to task engineers with preparing large and complex datasets for new use cases across the business.

Before the emergence of big data and cloud computing, data was largely regarded as an aggregation and management challenge. Data practitioners focused on database maintenance and operations, transactions were relatively linear, and data warehouse operations were often completed through lengthy and cumbersome batch processes that demanded more data practitioner time than acuity. From its roots in roles primarily focused on data wrangling, however, today data engineering has become a profession that encompasses data analytics, querying, behavior, processing, transparency, and maintaining reliability.[3]

This new importance has resulted in staggering growth in data engineering jobs. The 2021 Data Science Interview Report recorded 40% year-over-year growth in data engineering interviews, even as those for other data-related professions, such as data science, had leveled off.[4]

The degree to which data limitations hamper an organization's AI ambitions is well-understood in the C-suite. In our 2022 survey, 72% of executives agreed that data problems were the most likely issue to jeopardize their AI/ML goals.



As data volumes have grown, so have the number of business teams handling and interacting with data. This has widened the mandate of the data team from managing data pipelines, computation, and storage, to a performance-based role, where data is not only integrated, consolidated, and cleansed for use in analytics, but also made accessible, transparent, and optimized within the context of an organization's data ecosystem and digital estate.[5,6]

Amid these shifts, traditional foundational skills remain important for data engineers. These include a grounding in databases and programming; experience in data modeling, visualization, and integration; and knowledge of distributed computing, data lakes, and data streaming applications. But today's data engineers also find they need to know software engineering best practices, modern data integration methods, skills for working with data in real time (not just batched), programming languages such as Python and Scala, and platform cloud engineering. Data professionals must be equipped with the knowledge to manage and analyze complex datasets, understand the functionality of various platforms, the principles of data prioritization, and the limitations of data feasibility.

Srinivas Yadlapati, head of data at StockX, a leading online marketplace for trading and consuming current culture products including sneakers, apparel, accessories, electronics, collectibles, trading cards, and more, describes an increased need for knowledge of a range of programming languages, as well as a "convergence between data engineering and ML engineering," in terms of the tools and platforms used. Demand for soft skills is growing too. Being able to "solve problems, think outside the box and understand the bigger picture" is increasingly important for data engineers, says Venkatesh Kumar, director of data integrations at Apixio, a health-care AI provider.

As AI applications become more important to the business, data engineers must also master new types of data and programming languages; be able to use structured data, semi-structured data, and unstructured data from a variety of different sources; and manage a multitude of data pipelines. And as new regulations and requirements spring up around business uses of data and AI, data teams become responsible for architecting for governance, which requires a deep understanding of concerns such as data lineage, quality, and transparency.

New roles for data practitioners: Analytics engineers

Data engineer is not the only data role being invented – or reinvented – in the modern data organization. In the last few years, a new role, the analytics engineer, has emerged to bridge the space between data analyst and data engineer.

The adoption of new tools for data engineering – tools that bring the structure and rigor of software development to the management of data – as well as the advent of self-service data tools for the rest of the organization, have made it possible for former data analysts to move away from generating dashboards and reports and toward hands-on work with the data. Niklas Nordansjö, of digital energy provider Tibber, says, "with tools like we now use, you don't need much more technical knowledge than SQL to create the data models," paving the way for a new front-end role focused on "data modeling and understanding the data."

Databricks cofounder Reynold Xin explains, "Now I can turn virtually all my analysts into data engineers by giving them access to very structured business systems. They're calling them analytics engineers, which is a pretty cool new term. They're following data engineering principles, which means changes are captured, you can code review them, and you can do CI/CD, which brings rigor to the work and ensures that pipelines will run in a reliable fashion."

These analytics engineers can lighten the workload of data engineers and are becoming an increasingly essential interface between the business and its data. Drew Banin, cofounder at dbt Labs, says, "Analytics engineers are folks who understand the data domain. They really understand the business and the nature of the data that the business cares about, but they also understand the technology. So they're right in the middle and they can help translate the needs of the business into the data transformations that get applied in the data platform."

"When you start thinking about data as a product, you're compelled to work more like a product team. And then you can start to measure your success as a product team, prioritizing work that is much more strategic and high leverage."

Drew Banin, Cofounder, dbt Labs

The evolving data organization

The structure and positioning of the data organization within the enterprise is changing, as well. As access to – and responsibility for – data spreads across organizations, formerly centralized data team structures are in many cases becoming decentralized. Under these decentralized structures, each domain team may have its own processes for managing and analyzing its data assets – yet the data still demands collaboration and internal consistency.

Kumar claims that this democratization of data, together with low-code/no-code solutions, "is fundamentally changing the composition of data engineering teams," as employees from across the business need to understand data management and manipulation to do their jobs.

Valdar Tammik, head of data at autonomous delivery robotics company Starship Technologies, has noticed a similar trend, with data teams spending less time on building things in-house and traditional programming, and more on DevOps-style integration and system optimization. This, he says, "opens up avenues for more enablement work and work that is more heavily technical, as well as enabling others to work on data through data democratization."

A key organizational shift in many sectors has been recognizing the value and centrality of data to the business, and, consequently, creating more rigor and professionalism around how data applications are built and maintained. Known as "data as a product," this new emphasis has elevated the data team in many organizations, and it has changed their job descriptions as well. When they are seen as responsible for a business product, rather than just for maintenance of pipelines and tools, they're enabled to move away from IT-like support tasks and to lead on and prioritize work with substantial business impact.

Drew Banin, cofounder of dbt Labs, says, "When you start thinking about data as a product, you're compelled to work more like a product team. And then you can start to measure your success as a product team, prioritizing work that is much more strategic and high leverage." The shift to data as a product also gets data practitioners more deeply embedded in the business strategy – they must grow familiar with the business use cases they're supporting to properly shape and model data.

As such, today's data teams are working much more collaboratively, using platforms and tools whose design improves visibility, governance, and understanding around data for employees across the business. This paradigm is imposing new demands on data producers to ensure data reliability, completeness, and reusability, and providing opportunities to consider how data and data-led functions can integrate better with business teams.

Data ownership and responsibility

Treating data as a product transforms it into a more strategic tool. But it also poses questions about who is responsible for the data, how it fits in with the broader priorities and goals of the business, and the governance and management of data sharing. This "is exacerbating the issues of data security and provenance," says Kumar. "There is an inherent requirement from the data

engineering teams to deliver at light speed, but at the same time, to have to take all this into consideration."

In this environment, a high priority must also be placed on technical integration between data producers, software systems producing data, and data platforms, as not doing so can limit useful data insights and have an outsized impact on data quality. For many in the sector, this is a challenge that has yet to be overcome. "The fundamental issue hasn't really changed much," says Tammik. "There is limited technical integration between data producers and software systems and the platform. That means small changes in the software might have unintended consequences on pipelines and reporting," he adds.

Data contracts – specifications detailing aspects such as the content, format, and ownership of data – are an increasingly popular means of meeting these requirements.[7] Data contracts represent an agreement between data producers and data consumers, in which the producer commits to how it will provide data to the organization. Downstream users can consequently rely on receiving data of a certain quality, format, or timeliness, for example, and then confidently build those assumptions into their own data products. "Shifting ownership to the data producer with data contracts is very important in terms of collaborating around the data, trusting the data and building on each other's data," says Niklas Nordansjö, product manager at Tibber, a digital energy company.

Ensuring data quality for AI

Good models can't fix bad data. Data practitioners play a critical role in ensuring data quality. This imperative has come to the fore as generative AI, and particularly LLMs, show enormous potential to drive business value.

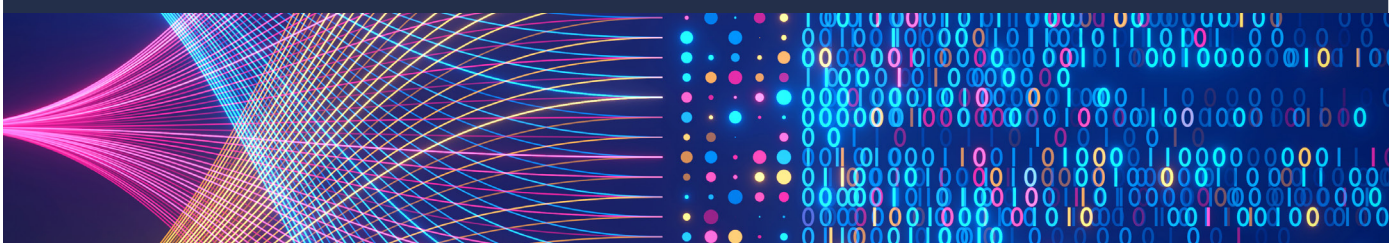
As AI applications begin to be widely deployed, data teams must have the skills and experience to understand the risks and limitations that come along with using LLMs and other generative models. These include biases in the underlying data sets, hallucinations (confident but fictitious AI outputs), and poor performance when not trained on context-specific data.

Retrieval-augmented generation (RAG), which grounds an AI model with a set of verifiable external knowledge sources, is emerging as a robust solution to some data quality concerns. Data teams using RAG supply their AI models with up-to-date and context-specific data, which may be unique or proprietary to their business, allowing for the production of higher-quality, more specialized, and – importantly – verifiable responses.

"Data engineers have an important role to fill here in terms of addressing the issue of hallucination," says Databricks cofounder Reynold Xin. "The way to avoid hallucination is to use retrieval-augmented generation. Instead of relying on the LLM for knowledge, RAG only uses the LLM to interpret the question, and then has the LLM go look at a source of truth that you deem correct. It could be a database, it could be a data lake, but you have data engineers involved to build that source of truth."

Ensuring the quality of the company's own data is of course a critical step. Modern data transformation tools and data platforms can help streamline and automate an organization's data wrangling, ensuring that its data is high quality and reliable. These solutions also give data practitioners clear visibility into the organization's data workflows, assisting them in making trusted data available to train AI models and to use in AI applications.

The need to customize, audit, and improve the quality of publicly available models, as well as to guarantee the quality of the organization's own data, underscores the importance of data practitioners in delivering successful AI results for any organization.



Partner perspective

Databricks

At Databricks, we believe that AI will eat all software. The implications of the recent disruptions in AI are far-reaching, including how they will change the way enterprises leverage data platforms. Currently organizations struggle with numerous challenges of developing AI applications on many data platforms. Barriers to getting value from data for AI applications range from technical skill gaps, poor performance and high costs, governance and privacy concerns (especially around lineage and security), data accuracy and accessibility, and decoupled data and LLM-tuning capabilities for developing custom generative AI applications.

Pioneered by Databricks, Data Intelligence Platforms address these challenges by leveraging generative AI to understand the unique semantics of an organization's data to analyze both the data and how it is used. The Databricks Data Intelligence Platform builds on the lakehouse foundation – an open, unified system to query and manage all data across the enterprise – and employs AI models to add new capabilities to all parts of the platform, automatically optimizing performance and managing infrastructure in ways unique to your business. It provides a unified query engine that spans ETL, SQL, machine learning, and BI. This means you can still do business intelligence and analytics with AI-enhanced tooling on the same data as new AI use cases that multiply the value of your enterprise data. Because it is built on open source and open standards, it is also simple to integrate with modern data and AI tools you have today, such as dbt-orchestrated SQL transformation workloads, or new technologies you may adopt tomorrow.

Databricks Data Intelligence Platform unifies data and governance to make the platform aware of company jargon, metrics, and semantics of the business. It enables advanced semantic search, better documentation, and AI assistant quality. Data teams can work with data in natural language to democratize insights for all users – from non-experts to data practitioners. It offers first-class support for AI workloads for generative AI and end-to-end AI applications to help deploy and manage models all the way through production.

In summary, the Databricks Data Intelligence Platform democratizes data and AI for the enterprise. By applying AI to every layer of the platform and using natural language, it radically simplifies the experience for everyone while achieving cost-efficiencies and automation that are specific to your business. Best of all, you own the data, giving you control and a strong competitive advantage.

Ken Wong
Senior Director of Product Management, Databricks

03

Data infrastructure and tooling for AI

One impetus for the changes in the data engineering profession has been the advanced tools and architectures now available for data management. Today's data practitioners have an increasing range of tools at their disposal to manage data, separate out data functions, process data in real time, and yield useful new insights.

These advanced data architectures and tools are both encouraging and requiring new ways of working for data practitioners. The modern suite of data tools allows data teams to professionalize, standardize, and simplify their work. Data practitioners can now access tools to support a standardized approach for building data products, much like those used by software developers: with built-in testing, continuous deployment, automated documentation, version control, lineage, discovery, and more.

Banin describes dbt Labs' work as "bringing software engineering best practices to data work." The company, he says, "gives data practitioners a standard way to model out data, what we call data transformation. It's taking all the source data that lands in your data platform and translating that data into a dataset that's ready for analytics, or BI, or machine learning."

Organizations also have options for modern data architectures and frameworks that enable them to make optimal use of their data assets. Data lakehouse and data mesh are likely to underpin many of tomorrow's data-driven business strategies, making understanding them a must for all savvy business leaders.

Centralization and data intelligence

For years, businesses have grappled with the best way to store and analyze growing volumes of data. From the integrity challenges posed by data from diverse sources,

> "Data platform is where we can enable all the data capabilities from an organization perspective. So we invested in a unified data platform that can support all these different workloads."
>
> Srinivas Yadlapati, Head of Data Engineering, StockX

emerged the idea of the data warehouse, a highly structured system for collecting that data and enforcing order. This was followed by the data lake, an unstructured way of collecting diverse types of data, and more recently the data lakehouse, which offers a unified system to query all an organization's data sources while also governing data workloads.[8] (See Figure 2.)

Yadlapati describes a strong desire for unification at StockX that led to the company's adoption of a lakehouse. "Data platform is where we can enable all the data capabilities from an organization perspective," he says. "I want a platform that can do real-time or batch processing. I want a platform that can do data warehousing. I want a platform that can do ML workloads. So we invested in a unified data platform that can support all these different workloads."

The lakehouse architecture is intended to reduce data silos that form when business teams store their data in separate warehouse ecosystems. Our report "Laying the foundation for data- and AI-led growth" found widespread uptake, with nearly three-quarters of surveyed organizations having adopted a lakehouse architecture – and almost all of the rest expect to do so in the next three years.[9] "The lakehouse gives you this capability to ingest data from all the different sources," says Reynold Xin, cofounder of Databricks. "It doesn't matter if it's unstructured, structured, or semi-structured. You can create this foundation for your large language models or a generative AI application to work against, and that gives you the best source of truth."

Data lakehouses also facilitate monitoring and analysis, can process all query languages, and help impose unified governance on diverse collections of data. Crucially, they also broaden the scope for collaboration. "One key component for us is allowing data analysts, machine learning engineers, and data engineers to work together on one platform, with real-time processing," says Yadlapati.

As AI emerges as a critical application for data, Databricks is focusing its efforts on a "data intelligence platform," a data lakehouse–based platform that brings together an organization's data and AI workloads on a single platform. Multiple teams can collaborate on the platform, enabling real-time data visibility across the organization.

According to Databricks, its data intelligence platform "understands the unique semantics of an organization's data that allows it to infuse AI in every part of the platform." It provides a tailored natural language interface for all users, followed by semantic cataloging and discovery, automated management, and optimization (in which data layout, partitioning, and indexing is customized based on usage), and enhanced governance and privacy capability (through automatic detection, classification, and protection of sensitive data). Finally, AI workload support is enhanced by allowing enterprise AI applications to connect to relevant business data.

The benefits of a unified data platform include streamlining of data management and workloads, real-time processing, and improved data warehousing. "A unified data platform can support lots of different workloads," says Yadlapati. "You don't want the data to move back and forth across data engineering, data warehousing, and ML processing."

Executed well, unified data platforms can deliver optimization for performance and storage; better indexing methodology; improved visibility, monitoring, security, permissions, and privacy; and a low total cost of ownership. A single platform also makes extraction,

> "The lakehouse gives you this capability to ingest data from all the different sources. It doesn't matter if it's unstructured, structured, or semi-structured.
>
> Reynold Xin, Cofounder and Chief Architect, Databricks

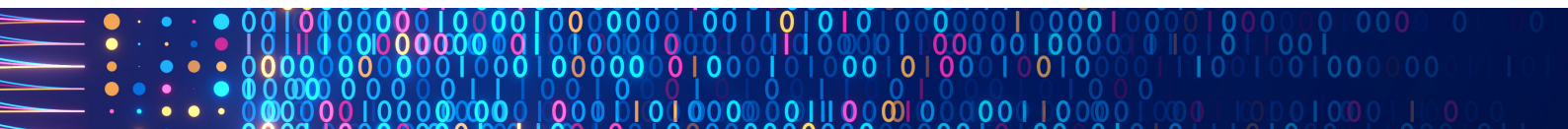
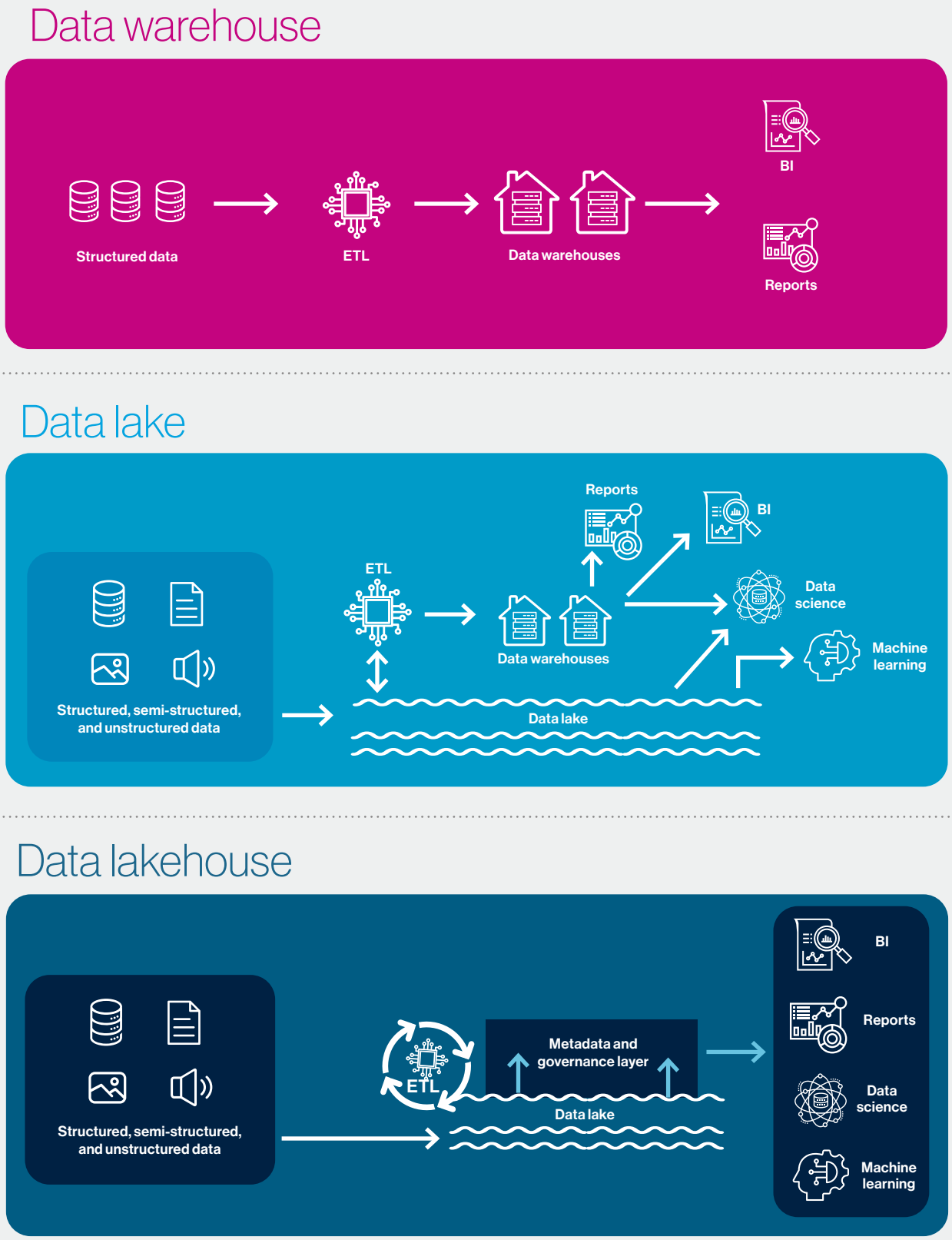


Figure 2: The evolution of data architecture options



Source: Compiled by MIT Technology Review Insights, based on data from "What is a lakehouse?" Databricks, 2024[10]

loading, and transformation (ELT) processes easier, allowing data to be collected in the lakehouse and then staged for use by other applications.

David Samet, director of technology at Fabuwood, a US-based cabinetry company, credits his company's unified data platform as a key differentiator in tackling supply chain challenges during the COVID-19 pandemic. "Because of our technology, tracking system, data collection, and analytics, all fully customized and developed in-house, we could ensure there were no surprises for our customers," he says.

The data mesh framework

As many organizations are finding success with a unified data platform, they are also finding that data is both produced and required by increasingly more different functions within the business. This trend is even more pronounced as the power of AI to transform the business becomes apparent.

Samet says Fabuwood is finding value in embedding data practitioners within business functions. "Connecting the users with the developers is the key," he says. "It changes the way data engineers think, act, and develop. Instead of having a team of data engineers, we're splitting the data engineers into different teams together with different data analysts based on the business domain."

Each part of the business may require the autonomy to build a data infrastructure that serves its needs, but these functions also require the support and governance of a centralized data management function. This need is driving a trend toward data platforms built within a framework called "data mesh."

Data mesh, a term first coined in 2019 by Zhamak Dehghani, is a framework that organizes data into domains, which are assigned to individual teams, each of whom have full and devolved ownership of the underlying

Four principles of data mesh

1

DATA DOMAINS

A data mesh is based on decentralized data ownership and architecture, with individual business domains owning their data and data products, along with the processes and pipelines that go with it.

2

DATA PRODUCTS

High-quality data is packaged by its owners in self-contained fashion, with the metadata, context, and tools that make it easy to use and create value from.

3

SELF-SERVE DATA PLATFORM

An accessible data platform allows business teams to use data and create new data products without centralized bottlenecks or dependence on data engineering teams.

4

FEDERATED COMPUTATIONAL GOVERNANCE

Each business team owns its own data, but shared governance creates interoperability and standardization, allowing for seamless organization-wide data use.

platform or data storage layer. Universal interoperability underpins each layer, such that the same syntax and data standards are applied in the interest of cross-domain collaboration, observability, governance, and standard-setting.[11,12]

This framework allows internal teams such as finance, product, and marketing to own their own data. However, rather than creating their own infrastructure from scratch, these teams are supported by a central data platform that helps govern and secure key data sources, while managing the processes for building data pipelines and assets that serve the entire business.

Domain-embedded data practitioners who work collaboratively with a centralized team can develop and manage their own pipelines and data products, which generally enables them to move faster and deliver more relevant insights. Once data is in the hands of data analysts closer to the business, greater cross-domain data analysis also becomes possible. This can open new analytical frontiers, raising the potential value and insights teams derive from data. In some cases, this can sharpen an organization's response to market conditions or make it more agile in the face of exogenous shocks.

In approaches to data organization like data mesh, where individual domain teams are responsible for their own data quality processes, individual data teams need structure and support from a centralized organization to be successful. One solution, championed by Tibber, is to use a "golden path" approach, where support is provided by the data team if other departments opt for the preferred platform, tools, and architecture. "If you need to do something with the data, ingest it or transform it, you have a golden path for that. And if you go by the golden path, then you will receive support. Which means we will validate the data, do tests on it, such that it ends up in a layer where we can access it and make use of it," says Nordansjö. That approach also makes the governance and data quality monitoring pieces of the data puzzle easier to manage. "We can control access to data, see the lineage and build trust, instead of having multiple places that you need to keep an eye on," he adds.

"Connecting the users with the developers is the key. It changes the way data engineers think, act, and develop."

David Samet, Director of Technology, Fabuwood

Partner perspective

dbt labs

In today's rapidly evolving business landscape, the integration of artificial intelligence (AI) into operations is no longer a luxury but an outright necessity. Companies worldwide are increasingly adopting AI initiatives to drive innovation, improve decision-making, enhance customer experiences, and deliver a competitive edge.

But for all its promise and investment, AI adoption is fraught with challenges, particularly in managing and leveraging the vast amounts of data that fuel these initiatives. The enthusiasm for incorporating AI into business operations is often dampened by the complexity of managing the underlying data infrastructure. A common challenge is the integration of large language models (LLMs) with cloud data platforms.

Without high-quality, continuously updated data connected to clear semantic definitions, businesses risk exposing inaccurate or outdated information to their LLM applications. Further, strong data governance practices are required to ensure that LLMs operate on data appropriately and that sensitive or regulated data is not misused or misconstrued by LLMs. Ultimately, LLMs are great at internalizing mountains of data and spitting out answers; if we prompt them with rich, high-quality inputs, we can expect to see high-quality outputs. Likewise, if we prompt with low-quality or non-compliant inputs, they may jeopardize business integrity and erode customer trust by outputting incorrect, invalid, or otherwise inappropriate outputs.

dbt Cloud is a powerful ally in this context, providing a suite of tools designed to maintain the integrity and trustworthiness of data powering AI initiatives:

• **Semantic layer implementation:** The creation of a semantic layer, which maps data to business concepts, ensures that the data exposed to AI models is accurate, relevant, and consistent. A semantic layer can significantly reduce the risk of hallucinations and inaccuracies.

• **Metadata framework:** A built-in metadata framework enables data to be enriched with a wealth of context and meaning, and it magnifies AI's ability to yield reliable answers to critical business questions.

• **Data quality with contracts:** Data contracts enforce clear definitions of data quality, structure, and relationships across teams. This ensures that only compliant data feeds into AI projects, even if that data crosses team boundaries, data stores, or domains

• **Version control and testing:** With built-in version control and testing capabilities, teams can track changes, test data models rigorously, and ensure that the data infrastructure remains stable and reliable.

• **Alerting and continuous integration:** Real-time alerting mechanisms to flag data quality issues, paired with continuous integration processes to catch issues before they hit production, ensure that data quality or model performance problems are promptly identified and addressed, preventing potential setbacks in AI initiatives.

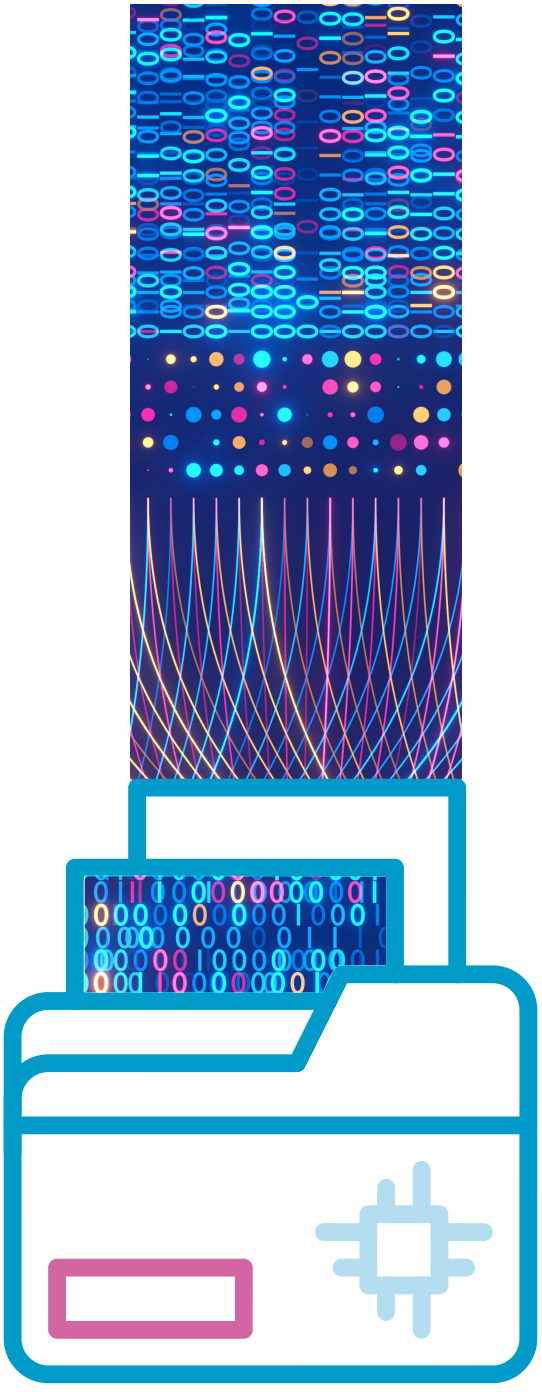
• **Accelerating AI projects:** AI initiatives can only move as fast as the development of the data that underlie them. dbt Cloud not only fortifies the data foundation of AI projects but also accelerates development and deployment by streamlining data transformation and modeling processes.

At dbt Labs, our mission is to empower data practitioners to safely create and disseminate organizational knowledge. These data practitioners will shape how AI is deployed in the enterprise and drive the strategy that leads to higher-quality results and designing data workflows that reduce risks in AI implementations. For them to be successful, we believe practitioners must adopt a structured and reliable approach to data management. By ensuring data integrity, fostering trust, and facilitating rapid development, dbt Cloud empowers businesses and data practitioners to leverage AI with confidence on top of their cloud data platforms.

Drew Banin, *Co-Founder, dbt Labs*

04

Data opportunities in enterprise AI



Organizations today face significant opportunities around data. They must plan for the changing role of data at the foundational level of the organization, while providing data practitioners with modern tools, delivering meaningful data access across the organization, and developing essential partnerships.

Enterprise AI is here today. Our 2022 survey of CIOs and technology leaders, "Becoming an AI-driven enterprise," found that 94% of organizations were already experimenting with or deploying AI in seven core business functions.[13] More than half of those surveyed expect AI use to be widespread or critical in these functions by 2025. The rapid emergence of generative AI only further underscores that seizing opportunities to invest in the data organization will create significant new discovery and analysis possibilities, with the potential for substantial value creation.

> "Being on a data team is one of the most challenging jobs, because you're spanning strategic prioritization and the needs of the business, but you're also responsible for execution."
>
> Drew Banin, Cofounder, dbt Labs

Supporting data practitioners for AI success

Today's data teams and practitioners can access a plethora of tools to help them deliver more incisive data-led insights. However, data practitioners also find they need to push the horizons of their own knowledge, as advances such as generative AI put ever more pressure on building and maintaining data infrastructure. The demands placed on them are increasing – from expanded skill sets to increased strategic business literacy, familiarity with new databases and tools, and responsibility for empowering data consumers across the business while ensuring sophisticated data governance and security.

Pressure on data practitioners is high, with many being asked to deliver more at a faster pace. "The demands placed on data engineers have skyrocketed," says Xin. "Data engineers are being asked to support a lot of high-priority activity, and I've never seen boardroom-level conversations being so aggressive and fast-moving."

The elevation of data discussions to the boardroom means engineers need to get closer to the business. Understanding the strategic "why" behind the technical "what" will help them deliver more timely and accurate insights. For many data practitioners, this is quite demanding. "Being on a data team is one of the most challenging jobs, because you're spanning strategic prioritization and the needs of the business, but you're also responsible for execution," says Banin.

Senior stakeholders in the business can, and should, help, by offering support to data practitioners. This could take the form of investment in technology and infrastructure that reduces the strain on data practitioners, including a modernized data stack. Samet, for example, says that Fabuwood supports a data culture by "investing a lot in modern data technology, where we can have the data stream flow much more easily, much smarter, from the beginning."

AI will of course play a part. AI-powered copilot tools for coding, for example, are already helping to optimize workflow at Starship, says Tammik. More sophisticated tools will be designed to bring the full power of AI to data engineering, unlocking unprecedented abilities to speed up data pipelines, optimize data workflows, automate scaling, and drive analytics.

Advanced automation will also facilitate data practitioners' work. More sophisticated automation could "enable systems and algorithms to become more data- and infrastructure-aware, which optimizes how data gets processed," says Kumar. Data pipelines driven by natural language processing (NLP) could also help create data models that are optimized as they evolve, Kumar adds, potentially creating "self-healing" data pipelines that can detect and recover from errors without human intervention.

Democratizing insight to empower everyone

As access to and responsibility for data spreads across the organization, investing in data literacy for everyone is important. Every employee who engages with the data platform needs to understand the platform's functionality, data prioritization, and data feasibility limitations.

Functional areas of the business can put domain-area knowledge into practice if they understand how to access, manipulate, and interpret the data they have access to. "To have a data culture where decisions are being made out of data and not out of gut feeling," says Samet, "we need to be able to give the user the data at their fingertips." Working with embedded data practitioners who can help teams develop and manage their own pipelines and data products can help business teams gain greater insights and solve their own problems. "Once users see, 'Hey, data is actually showing me something, data gives me insights that I didn't know before,'" Samet adds, "they instantly want to see a lot more, and they want to make smarter decisions based on data."

Enabling this requires that businesses invest in support for less technical users. These might include simplified tools for working with data, such as low-code toolkits for analysts, which allow non-experts to model and access data without sophisticated knowledge of complex data engineering practices. Alternatively, they could be as expansive as a data intelligence platform that uses AI, grounded in deep contextual knowledge of the business, to answer natural-language data questions posed by any user.

Even then, however, not all employees will feel equipped to work with data. "Democratization opens up challenges," says Tammik. "A lot of people are not used to working with data or designing data sets, which means all sorts of suboptimal designs could pop up." So while democratization of data and AI is a net positive, ensuring that people are capable of working with data and AI effectively is still a challenge. This takes tooling with guidance and guardrails that encourages collaboration in a governed way.

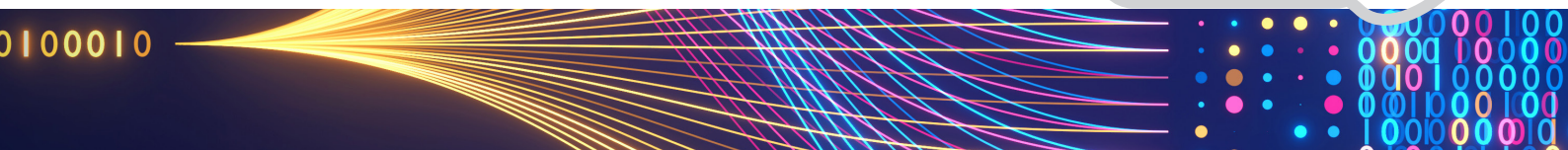
Ensuring governance and oversight for data and AI

Opening data access across the organization – or even to external partners – underscores the need for robust governance and oversight. As LLMs and other generative AI applications cause data to spread even more broadly across organizations, and more stakeholders are enabled to make use of data and insights, governance will become ever more essential. Some organizations find that these oversight requirements are best merged and addressed via a unified platform. Xin says, "Security and data governance is now increasingly more critical for a lot of reasons. A single unified governance platform means you can think about things holistically; you know whether a model is trained or inferred on specific datasets."

Kumar agrees about the governance benefits of a centralized data platform, saying that a recent move to a unified platform enhanced Apixio's efficiency and understanding of its data provenance, which improved quality for both consumers and internal models. Nordansjö adds, "By having everything in the same place, we can control access to data, we can see lineage, and we can build trust in our data, instead of having multiple places you need to monitor. We have a better overview of the data when it comes to governance and data quality and such."

Another opportunity to help address the governance challenges presented by LLMs and generative AI is to holistically integrate AI use cases into data workflows that are governed, tested, and secured according to industry best practices. To do this, organizations should ensure that any data-based AI applications sit on top of a codebase following best practices such as version control, testing, and data access controls.

> "To have a data culture where decisions are being made out of data and not out of gut feeling, we need to be able to give the user the data at their fingertips."
>
> David Samet, Director of Technology, Fabuwood



05

Conclusion

Modern data tools and technologies can deliver business insight that creates efficiencies, maximizes value, services customers, and increases organizational agility. But the road to these data-driven riches is scattered with obstacles. How data practitioners and their organizations respond to these challenges will largely determine their success in this era of data proliferation – and their ability to participate in the coming AI revolution.

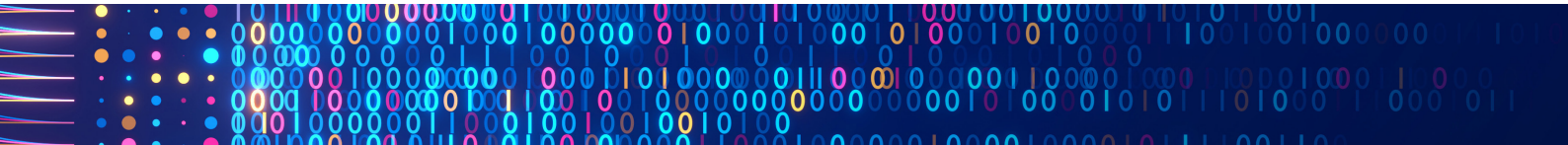
Organizations must consider how their data organization is structured and their investments in their data teams, making choices that will enable their data practitioners to set up the business for success. They need to make infrastructure and architectural choices that will be the foundations of the data-driven future, and also make the cultural and values shifts around how their organization values data.

What does all this mean for data practitioners? As their organizational importance increases, they will need to be more knowledgeable, agile, and able to collaborate. They will also need to be well-versed in the strategic direction of business verticals – and the goals of the organization as a whole – to be able to serve their needs and those of their colleagues more effectively.

Xin underscores the critical role of data practitioners in their organizations' AI ambitions. "Many people, including most data engineers, feel like AI is remote," he says. "But the reality is that what you do as a data engineer lays the foundation for AI. It's absolutely necessary, and you have power to make or break all of these AI applications. As a matter of fact, you probably have more power than the people who are actually building the AI models, because the models are now built to be more and more tolerant, but it's the data that's making all the difference."

Business leaders also have a critical role to play in bringing their organizations into the data and AI future. This is a key moment for them to make the right organizational shifts and critical data investments to enable their companies' continuing relevance. Individual stakeholders have their part to play as well. "For this to really go well," says Banin, "two things need to happen. One is the data team needs to get closer to the business and understand it. But two is that, as a stakeholder in the business, it's increasingly becoming your responsibility to become data literate, and understand how to interface with data, and ask good questions about data."

> "The reality is that what you do as a data engineer lays the foundation for AI. It's absolutely necessary, and you have power to make or break all of these AI applications."
>
> Reynold Xin, Cofounder and Chief Architect, Databricks



About MIT Technology Review Insights

MIT Technology Review Insights is the custom publishing division of *MIT Technology Review,* the world's longest-running technology magazine, backed by the world's foremost technology institution—producing live events and research on the leading technology and business challenges of the day. Insights conducts qualitative and quantitative research and analysis in the U.S. and abroad and publishes a wide variety of content, including articles, reports, infographics, videos, and podcasts. And through its growing MIT Technology Review Insights Global Panel, Insights has unparalleled access to senior-level executives, innovators, and thought leaders worldwide for surveys and in-depth interviews.

About Databricks

Databricks is the Data and AI company. More than 10,000 organizations worldwide — including Comcast, Condé Nast, Grammarly, and over 50% of the Fortune 500 — rely on the Databricks Data Intelligence Platform to unify and democratize data, analytics, and AI. Databricks is headquartered in San Francisco, with offices around the globe, and was founded by the original creators of Lakehouse, Apache Spark™, Delta Lake, and MLflow. To learn more, follow Databricks on **LinkedIn**, **X**, and **Facebook**.

About dbt Labs

Since 2016, **dbt Labs** has been on a mission to help analysts create and disseminate organizational knowledge. dbt Labs pioneered the practice of analytics engineering, built the primary tool in the analytics engineering toolbox, and has been fortunate enough to see a fantastic community coalesce to help push the boundaries of the analytics engineering workflow. Today there are 30,000 companies using dbt every week and over 4,100 dbt Cloud customers.

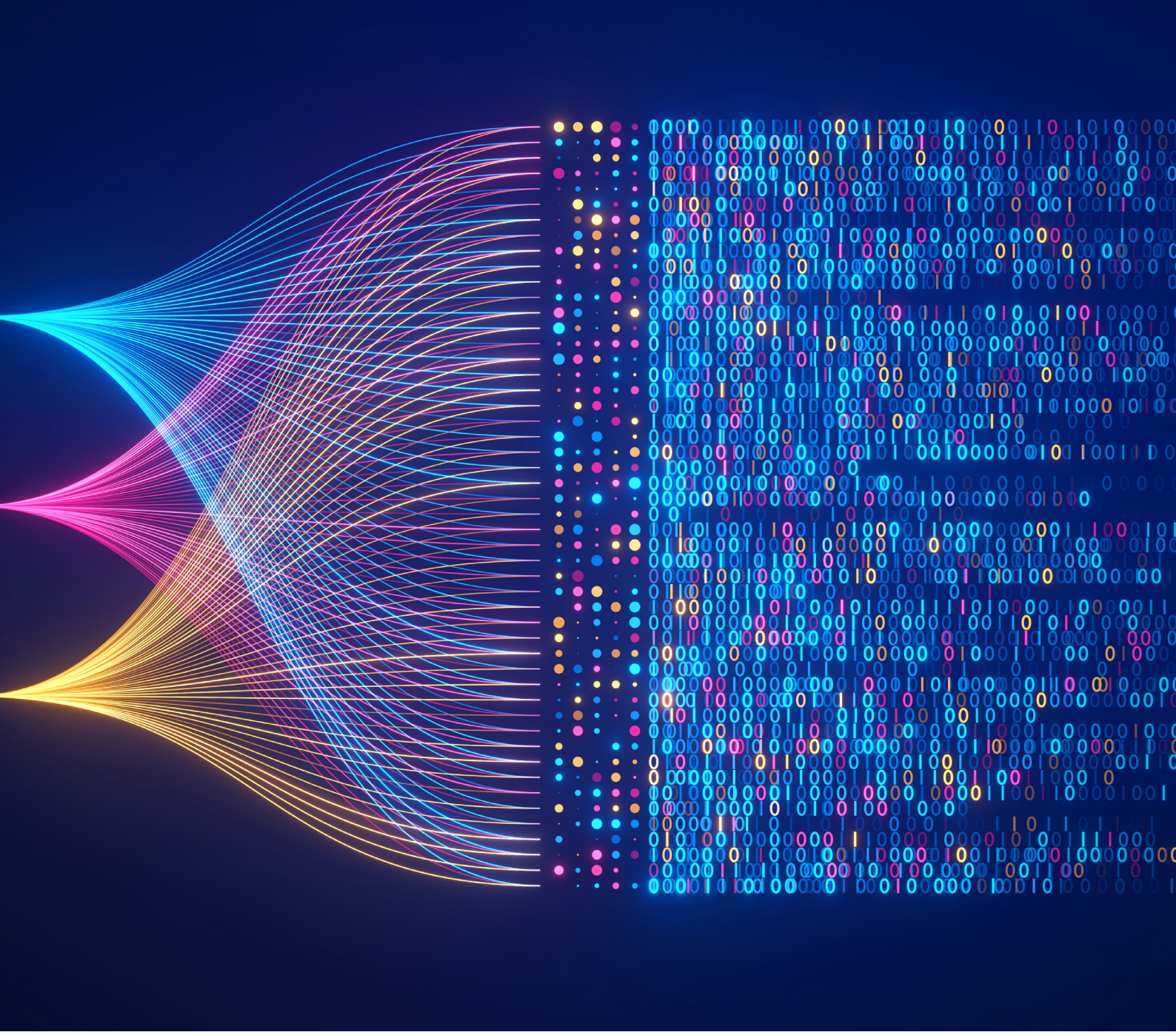
databricks × dbt Labs

Endnotes

1. "Becoming an AI-driven enterprise," MIT Technology Review Insights, September 20, 2022, https://www.technologyreview.com/2022/09/20/1059630/cio-vision-2025-bridging-the-gap-between-bi-and-ai/.
2. "Becoming an AI-driven enterprise," MIT Technology Review Insights, September 20, 2022, https://www.technologyreview.com/2022/09/20/1059630/cio-vision-2025-bridging-the-gap-between-bi-and-ai/.
3. Tony Baer, "The evolving role of the data engineer," Silicon Angle, February 10, 2023, https://siliconangle.com/2023/02/10/evolving-role-data-engineer/.
4. "The 2021 Data Science Interview Report," Interview Query, May 3, 2023, https://www.interviewquery.com/p/data-science-interview-report.
5. Lior Gavish, "The future of data engineering as a data engineer," Monte Carlo, January 23, 2024, https://www.montecarlodata.com/blog-the-future-of-the-data-engineer/.
6. Ben Lutkevich and Jack Vaughan, "Definition: data engineer," TechTarget, March 2021, https://www.techtarget.com/searchdatamanagement/definition/data-engineer.
7. Michael Segner, "Data contracts – everything you need to know," Monte Carlo, December 8, 2022, https://www.montecarlodata.com/blog-data-contracts-explained/.
8. "Data lakehouse," Databricks, https://www.databricks.com/glossary/data-lakehouse.
9. "Laying the foundation for data- and AI-led growth," MIT Technology Review Insights, October 5, 2023, https://www.technologyreview.com/2023/10/05/1080618/laying-the-foundation-for-data-and-ai-led-growth/.
10. Ben Lorica, Michael Armbrust, Reynold Xin, Matei Zaharia, and Ali Ghodsi, "What is a lakehouse?" Databricks, January 30, 2020, https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html.
11. "Demystifying data mesh," McKinsey & Company, June 8, 2023, https://www.mckinsey.com/capabilities/quantumblack/our-insights/demystifying-data-mesh.
12. Idowu Odesanmi, "What is a data mesh architecture?" SingleStore, January 12, 2023, https://www.singlestore.com/blog/what-is-a-data-mesh-architecture/.
13. "Becoming an AI-driven enterprise," MIT Technology Review Insights, September 20, 2022, https://www.technologyreview.com/2022/09/20/1059630/cio-vision-2025-bridging-the-gap-between-bi-and-ai/.

Illustrations

Cover art by Adobe Stock. Spot illustrations assembled by Chandra Tallman Design with icons provided by The Noun Project and Adobe Stock.



MIT Technology Review Insights

www.technologyreview.com

insights@technologyreview.com